DOCUMENT RESUME

ED 175 941 TM 009 680

AUTHOR Veloski, Jon
TITLE Prediction of Pass/Fail on a Certifying Examination Using Discriminant Analysis with Cross Validation.
PUB DATE Apr 79
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Discriminant Analysis; Higher Education; *Mastery Tests; *Medical Students; *Predictive Validity; Remedial Programs; *Science Tests; Student Certification
IDENTIFIERS *Cross Validation

ABSTRACT

Discriminant analysis was used to predict the performance of medical students on a certifying examination, uisng available measures approximately seven months in advance of the examination. The purpose was to identify those students having the greatest chance of failing in order to provide them with remedial help. The linear discriminant function was derive based upon the historical performance of three classes of students, and the model was cross-validated on two subsequent classes. Although the criterion established for determining the success of the predictive model was met in the first validation, it was not met in the second. (Recommendations for revisions of the model and the importance of cross validation are discussed. (Author/MH)

Prediction of Pass/Fail on a Certifying Examination Using Discriminant Analysis with Cross Validation*

Jon Veloski
Jefferson Medical College
Thomas Jefferson University

In order to predict the performance of medical students on a certifying examination using available measures approximately seven months in advance of the examination, discriminant analysis was used. The linear discriminant function was derived based upon the historical performance of three classes of students (n=638) and the model was cross validated on two subsequent classes (n=217 n=209). Although the criterion established for determining the success of the predictive model was met in the first validation, it was not met in the second. Recommendations for revisions of the model and the importance of cross validation are discussed.

* Presented at a round table discussion at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1979.

Prediction of Pass/Fail on a Certifying Examination
Using Discriminant Analysis with Cross Validation

Jon Veloski
Chief, Evaluation Section
Office of Medical Education
Jefferson Medical College
Thomas Jefferson University

The ability to predict behavior is one of the more important aims of research in the behavioral sciences. The performance of students on written examinations has been predicted using the techniques of multiple linear regression.[1] Other multivariate techniques, however, provide results that are comparable for certain purposes, and some of these techniques may provide other useful benefits that are not readily obtainable when multiple linear regression is used. The present study involved the use of discriminant analysis to predict performance on a written examination, and it illustrates some of the potential secondary benefits of this approach.

Since 1970, the medical college has used multiple linear regression to predict the performance of second-year medical students on a standardized certifying examination which is regularly administered at the end of the sophomore year. The predictions from available measures have been generated seven months prior to the examination data, and those students predicted to have difficulty in passing the certifying examination have been counseled accordingly, while special attention has been directed to the students who are expected to do well on the examination.

The certifying examination is administered during two consecutive days; it contains over 1000 multiple-choice questions in seven basic science disciplines: anatomy, physiology, biochemistry, pathology, pharmacology, microbiology and behavioral science. A total score and each of the seven subtest scores are reported on a scale of 200 to 800. The examination is constructed and aministered by a national testing board to students in many medical schools throughout the country, and the scores are standardized to a national mean of 500 and standard deviation of 100. The "passing" score on the certifying examination is 380, and in earlier years about 97 percent of the students at this medical college achieved this score or better. Satisfactory performance on the examination requires only a passing total score; an examinee may fail one or more of the subtests but achieve an overall passing score as a result of strengths in other areas of the examination.

The purpose of the annual prediction effort at the medical college is to identify only those students who might be expected to have the greatest change of failing the examination and to provide those "academically marginal" students with remedial assistance in order to lessen their chances of failing. The experience of seven years has shown that the identification of the approximated 30 students of a total class of 220, having the lowest scores predicted by the multiple linear regression equation assured selection of at least 80 percent of those who eventually

fail. In November, 1976, for example, 28 students were selected by the prediction model. In June, 1977, 12 of the 220 students failed the examinations, and of these, 10 had been included in the predicted group.

Although the prediction system based upon the regression model has proved satisfactory over the years, several problems have been identified. One is that the predictions provide more specificity than is actually required by the faculty and administration, who are not generally concerned with a student's relative predicted score but want to know his or her likelihood of passing or failing. A second undesirable characteristic of the regression approach is the measure by which its accuracy is gauged: the multiple correlation or its square, the percentage of variance. Experience has shown that persons unfamiliar with elementary statistics may have difficulty interpreting this index. A third difficulty arises when one tries to interpret the validation of a set of predictions. The faculty and administration are not necessarily interested in the validity coefficient or the scatter plot of predicted and obtained scores. They want to know the number of students who were predicted to fail, and did not; and the number who were not predicted to fail, and did. A desirable predictive technique would therefore be not only reliable, but also understandable and of course, economical.

Given the purposes and uses of the predictions and the less than optimal aspects of the multiple linear regression technique, discriminant analysis seemed to be a reasonable alternative. The present study was not intended to be a formal comparison of the regression technique and the discriminant analysis. Rather, it was intended to be an exploratory investigation to determine if discriminant analysis would provide reasonable results when applied to the same student data on which the regression model had been used, and to learn whether the classification information provided by the discriminant analysis would alleviate some of the problems with multiple regression discussed in the introduction. At the beginning of the study these hypotheses were advanced:

1. Discriminant analysis will provide predictions for a pass/fail classification based upon a dichotomy of the certification examination score. These predictions will be less precise than those obtained by multiple regression.

2. The results of the discriminant analysis, although less precise, will be easier to report and interpret in review of the fact that the intended use is to predict pass/fail on the certifying examination.

It was decided a priori that, in order to provide usable results, the discriminant analysis would necessarily predict at least 80% of the actual failures in a class while selecting not more than 30 students as expected to fail. For example, if in a given class the model predicted 30 of 215 students to fail, and ten from the entire class did fail, at least eight of the ten must have been included in the predicted group in order for the prediction technique to be judged effective.

As recommended by Cooley and Lohnes it was also decided that a proper validation of the method of discriminant analysis must confirm the stability of the predictions on a group or groups other than the sample used for the initial generation of the discriminant model.[2] This requirement suggested a need for several years' data in order to test the new method.

Method

In order to validate the discriminant analysis approach, it was decided that the prediction would be simulated retrospectively by using data on five classes of medical students. The three earliest classes were used for the calibration sample, and the two later classes each constituted an independent validation sample.

The subjects included all new matriculants at the college between the years 1972 and 1976, except those s udents who withdrew, took leaves of absence from the school, or dropped back after failing courses prior to the certification examination administered at the end of the second year. For this study the subjects were separated into the three samples:

(a) the calibration sample, those who entered between the years 1972 and 1974,

(b) the first validation sample, those who entered in 1975, and

(c) the second validation sample, those who entered in 1976.

Diagram I shows the approximate chronology of the discriminating variables and the certifying examination. The variables: Medical College Admissions Test-Science subtest, Function and Structure, and Pathology were selected as discriminating variables for this study based upon previous studies conducted to select variables for use in the multiple linear regression model. These variables have shown useful correlations with the criterion certifying examination while having lower multicolinearity than other measures.

The Medical College Admissions Test (MCAT) Science score is derived from a standardized examination administered nationwide to medical school applicants. It is reported on the scale of 200 to 800 and is designed to measure aptitude and achievement in the basic science disciplines of biology, physics, and chemistry. Function and Structure is a grade based upon internal objective examinations administered at the medical college during the first year. This grade is reported on a 0 to 100 scale, has a standard error of measurement of 2 and is intended to reflect achievement of course objectives involving knowledge of human anatomy and physiology. The pathology grade is based upon objective examinations administered at the medical college during the beginning of the second year. This grade is also reported on a 0 to 100 scale and has a standard error of measurement of 1.

The data for discriminating variables and the scores on the certifying examination were obtained from the data base of a longitudinal study of the medical school classes. Using a large general-purpose computer, the analyses were performed using the Statistical Analysis System (SAS76).[3] This particular program was selected because of its ability to handle easily the three samples of students and specifically to perform a discriminant analysis on a calibration sample and to apply the obtained model to other subjects for validation. In addition, SAS76 contains a powerful plotting procedure which was selected to perform post hoc plots of the discriminating variables and the subject classifications. Huberty[4] states that the Statistical Package for the Social Sciences (SPSS) contains a most complete discriminant analysis program. The stepwise analysis feature of that program might have been useful if prior information to guide the selection of discriminating variable had not been available.

In order to establish classifications for discriminant analysis, the criterion score on the certifying examination was dichotomized for the calibration and validation samples, using a cutoff of greater than or equal to 380 assigned "pass", and less than 380 assigned "fail". Discriminant analysis was performed on the calibration sample. The prior probabilities for the classification analysis were established at levels proportionate to the distribution of the pass/fail criterion in the calibration sample. The classification model derived from this sample was subsequently applied to the first validation sample.

The results of the first trial suggested that the pass/fail threshold might be raised to 410 in order to take into account the error of measurement of the certifying examination. The calibration sample was analyzed again using the procedures outlined above and predictions generated for the first validation sample.

A third trial was attempted. Each of the subjects in the calibration and validation samples was assigned to one of six categories based upon his/her criterion score. For these categories intervals of thirty score units (approximately one standard error of measurement) were used up to 469 and two larger intervals at the upper end of the scores (470 to 599, and 600 up). Again, the calibration sample was analyzed using the procedures outlined previously. The single validation sample was classified based upon the new model for six categories of the criterion. Finally, the second validation was classified using the new model.

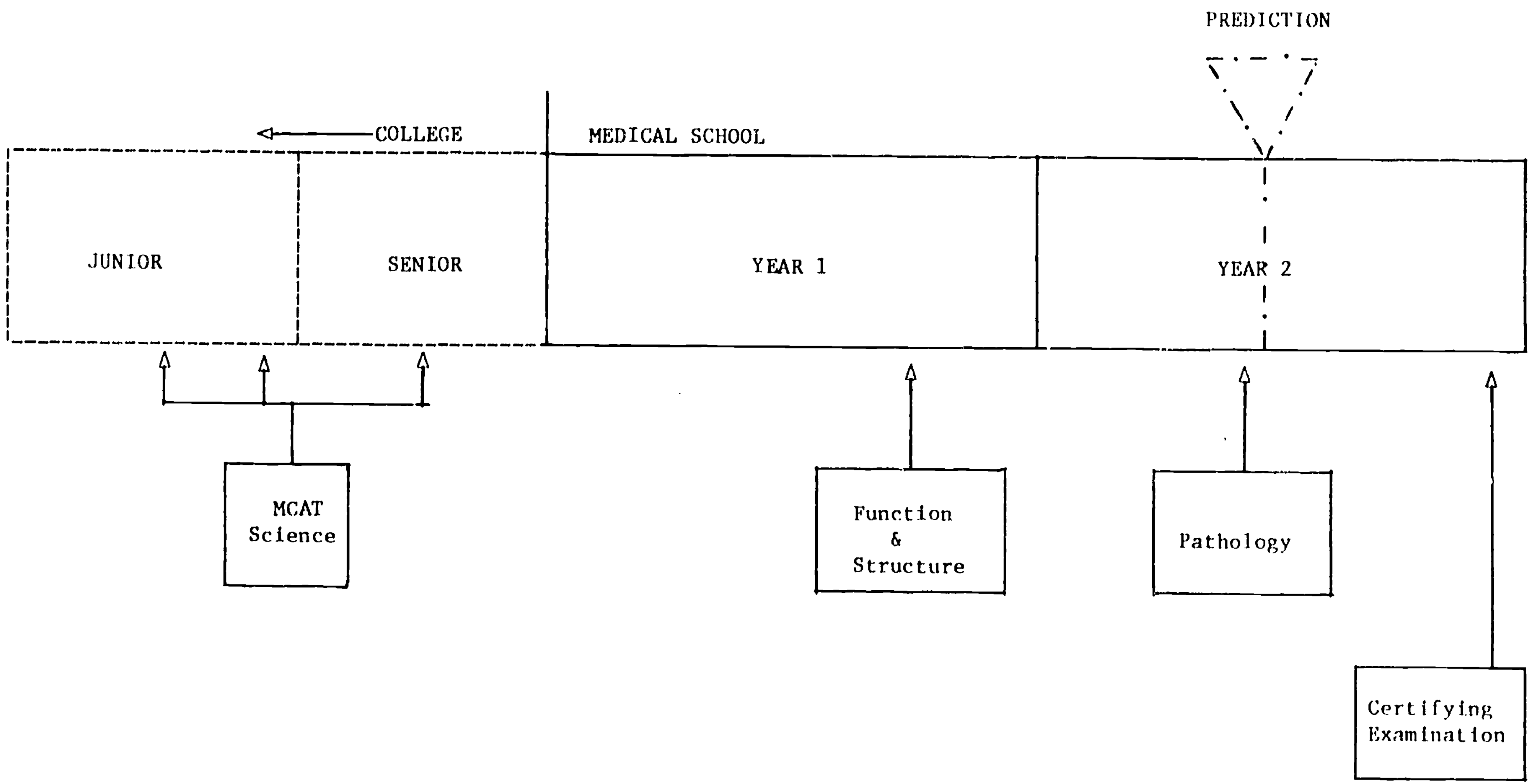
Based upon the data derived from the third model, reports of group centroids were prepared using SAS76, and these were manually superimposed on scatter diagrams using symbols to denote the various score intervals (Diagrams II and III).

Results

Table I shows the descriptive statistics for the three samples of students. Each sample nearly represents the total population of Jefferson

DIAGRAM I

CHRONOLOGY OF THE PREDICTIONS



students writing the examination in given years, so there is no concern for the representativeness of the samples used for this study.

The results of the initial discriminant analysis were surprising. Using a cutoff of 380 for pass/fail, the classification procedure applied to the calibration sample identified only 51% of the actual failures, although it did this by selecting only 5% of the sample as potential failures. When the classification rules derived from this analysis were applied to the validation sample, only 20% of the actual failures were successfully classified.

The outcome of the second trial is shown in Table II. These results show little gain in predictive accuracy after the pass/fail threshold was increased to 410.

The third trial provided results that met the standards set forth at the beginning of the study. Table III is the summary of classification successes and failures for the calibration sample. It can be observed that 66% of the 638 students were classified properly. In this analysis the four predicted categories for scores less than 470 comprised 13% of the total sample, and these accounted for 82% of the actual failures.

Of greater interest is the result obtained when the obtained classification model was applied to the first validation sample, as shown in Table IV. Of the 217 students, 58% were properly classified. All but one of the actual failures were predicted in the four groups below 470 which would have identified 11% of the students as potential failures. These potential failures accounted for 91% of the actual failures.

The values of the group centroids for the calibration sample are shown in Table V. In Diagram II the group centroids for the calibration sample are superimposed on the scatter plot of the MCAT Science scores (horizontal) versus the Function and Structure grades (vertical) for the validation sample. Diagram III shows the group centroids and plots for Function and Structure versus the Pathology grades. The centroids and scatter plots present a variety of data which enable the reader to make interpretations of the discriminant analysis. Table V suggests that the MCAT science is an effective discriminator among the upper four groups and between these and the bottom two, but contribute nothing to the discrimination between the lowest two groups. This observation is confirmed by Diagram II where the centroids are plotted.

It can be observed that Function and Structure grade is the most consistent discriminator of the six groups. The plot shown in Diagram III helps to amplify the discrimination, but it also calls attention to a relative discontinuity of the Function and Structure grade discrimination between the groups (?) and (*), which reflect criterion scores of 410 to 439, and 440 to 469, respectively. It is important to observe that the discontinuous region of the plot appears to coincide with the outer boundary of the failure (F) group. Students in the validation sample with Function and Structure grades above 77 did not fail, and students with grades in pathology above 80 never failed.

Table VI shows the results of the application of the classification model to the second validation.

<u>Discussicn</u>

Although the results obtained in the first two trails were indeed disappointing, the third attempt produced acceptable classifications. These findings suggest that a dichotomized classification variable based upon a pass/fail threshold of a continuous variable will not permit a useful discriminant analysis. Even when the threshold was increased, little improvement was observed. When the number of classifications was expanded to six, the discriminant analysis generated predictions which met the standards set in the study objectives.

The results show that for the first validation sample the failure predictions generated by the discriminant analysis model were as good as, if not better than those generated by the multiple linear regression model. Selecting comparable numbers of students "at risk", the regression model identified 9 of the 11 failures, while the discriminant analysis model identified 10. It should be noted that a followup indicated that the one failing student not predicted by the discriminant analysis was not one of the two failures missed by the linear regression. This finding should be interpreted with caution since the regression model used for the 1975 sample (the validation sample) was derived from only the previous years' data and not three years' data, as was the discriminant analysis predictive model in this study. The results of the second cross validation were indeed disappointing. The percentage of the failures included in the "at risk" group did not meet the standards set forth at the beginning of the experiment. The descriptive statistics of the independent variables shown in Table I do not show a change in the second validation sample.

One of the more useful findings of this study might be the utility of the tabular and graphic presentations of data that are readily available for discriminant analysis results. The classification table of successes and failures and the scatter plot of independent variables could be generated from multiple regression results, with some effort, but they are a simple by-product of discriminant analysis. The classification tables provide concise summaries of the validity of the model when applied to subsequent samples, and the scatter plots with symbolic identification of membership in each criterion class provide an approximation of a three-dimensional plot.

There are a number of assumptions in this study that have been left untested but that might deserve further work. The fact that the students who are predicted to fail are counseled may invalidate the predictive models in the future, although seven years' experience does not support this. The effects of the varying means and standard deviations of the three samples reported in Table I were not examined, and these may have a significant effect on the predictions. According to Huberty, a third area that deserves further work is the differing

cost of certain types of prediction errors have not been given adequate treatment.[4] For example, the error which predicts students to pass and who actually fail may be more costly than the error which predicts students to fail, and who eventually pass.

Conclusions

The findings reported show that discriminant analysis can be used to generate predictions of failures in the population of students at this school. The use of two independent validation samples demonstrates that the findings are generalizable to one subsequent group, but the application of the classification model to a second validation sample does raise questions about the generalizability of the model.

REFERENCES

1. Best, W.R., et al, Multivariate Predictors in Selecting Medical Students, Journal of Medical Education, 46:42-50, 1971

2. Cooley, W.W. & Lohnes, P.R., Multivariate Data Analysis, New York, Wiley, 1971

3. Barr, A.J., et al, A User's Guide to SAS76, Raleigh, North Carolina State University Student Supply Store, 1976

4. Huberty, C.J., Discriminant Analysis, Review of Educational Research 1976, 45:543-598

TABLE I

Means for Standard Deviations

for the Three Sample Groups

Sample	n	Means: MCAT Science	Means: Function and Structure	Means: Pathology	Standard Deviations: Science	Standard Deviations: Function and Structure	Standard Deviations: Pathology
Calibration (1972-74)	638	617	83	82	67	6	5
First Validation (1975)	217	634	82	83	76	5	5
Second Validation (1976)	209	623	82	83	76	6	6

TABLE II

Classification Results of

Discriminant Analysis for the Second Trial

(Pass: > 410 Fail: ≤ 410)

Calibration Sample
(n=638)

Actual Results	Predictions: Pass	Predictions: Fail
Pass	585	8
Fail	29	16

First
Validation Sample
(n=217)

Actual Results	Predictions: Pass	Predictions: Fail
Pass	188	2
Fail	22	5

TABLE III

Classification Results of

Discriminant Analysis for the Third Trial

Calibration Sample (n=638)

	Prediction					
	- - - - - - At Risk - - - - - - - -					
	Fail	380-409	410-439	440-469	470-599	600-up
Actual Result						
600 up					50	76
470-599		2	2	10	314	29
440-469	1	3	2	14	53	
410-439	2	5	4	10	15	1
380-409	2	7	0	6	8	
Fail	7	4	2	5	4	

O = correct classification

Correct Classification = $\frac{422}{638}$ = 66%

Failures included in "at risk" group = $\frac{18}{22}$ = 82%

TABLE IV

Classification Results

for First Cross Validation

First Validation Sample (n=217)

	Prediction					
	- - - - - - - At Risk - - - - - - - -					
	Fail	380-409	410-439	440-469	470-599	600 up
Actual Result						
600 up					9	31
470-599				2	89	16
440-469			2	2	20	
410-439			1	2	16	
380-409	1			3	12	
Fail	2	3	1	4	1	

Correct Classification = $\frac{125}{217}$ = 58%

Failure included in "at risk" = $\frac{10}{11}$ = 91%

TABLE V

Means and Standard Deviations

for Centroids of Discriminating Variables

Category based upon certifying examination	Means			Standard Deviations		
	MCAT Science	Function and Structure	Pathology	MCAT Science	Function and Structure	Pathology
(.) 600 and up	656	89	88	58	4	4
(+) 470 to 599	620	83	82	55	5	4
(*) 440 to 469	599	78	78	66	5	4
(?) 410 to 439	582	77	76	82	5	4
(=) 380 to 409	536	76	76	70	5	3
(F) Fail	538	74	76	92	5	4

TABLE VI

Classification Results

For Second Cross Validation

Second Validation Sample (n=209)

	Prediction					
	- - - - - - - - At Risk	- - - - - - - - -				
	Fail	380-409	410-439	440-469	470-599	600 up
Actual Result						
600 up				1	6	25
470-599				4	67	9
440-469				4	20	
410-439	1	2	1	2	19	
380-409	0	1	1	4	11	
Fail	4	3	4	9	11	

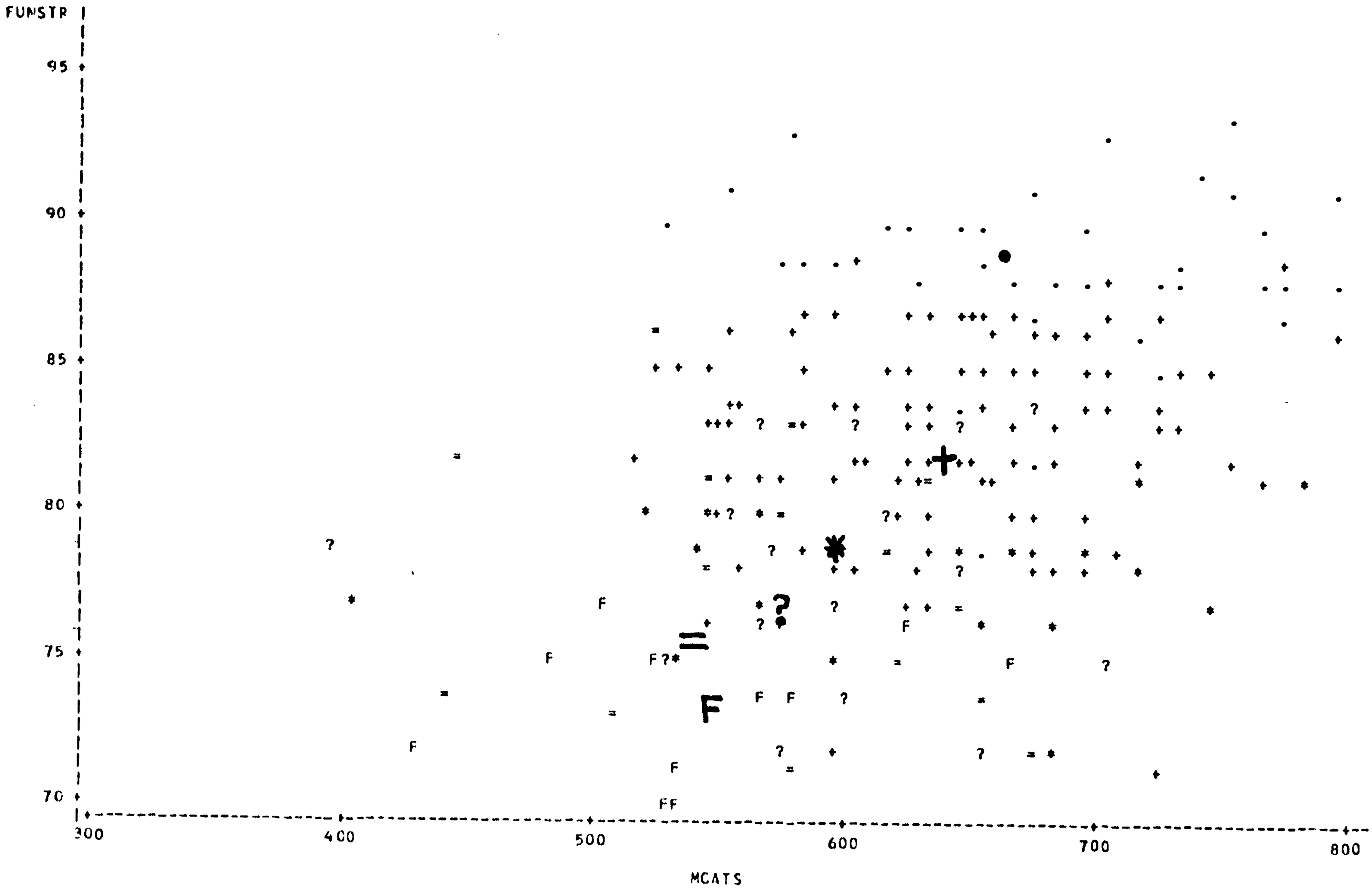
Correct Classification = $\frac{102}{209}$ = 49%

Failures included in "at risk" group = $\frac{20}{31}$ = 65%

DISCRIMINANT FUNCTION ANALYSIS PASS AND FAIL PART I OF NBME
BASED UPON ENTERING CLASSES OF 1972 THROUGH 74
VALIDATED ON CLASS ENTERING 1975
PREDICTIONS GENERATED FOR CLASS ENTERING 1976
FAIL(F) 380-409(=) 410-439(?) 440-469(*) 470-599(+) 600- (.)
PREPARED BY J.V. 11/77

Diagram II

PLOT OF MCATS*FUNSTR LEGEND: SYMBOL IS VALUE OF RESULT



DISCRIMINANT FUNCTION ANALYSIS PASS AND FAIL PART I OF NBME
BASED UPON ENTERING CLASSES OF 1972 THROUGH 74
VALIDATED ON CLASS ENTERING 1975
PREDICTIONS GENERATED FOR CLASS ENTERING 1976
FAIL(F) 380-409(=) 410-439(?) 440-469(*) 470-599(+) 600- (.)
PREPARED BY J.V. 11/77

Diagram III

PLOT OF FUNSTR*PATH LEGEND: SYMBOL IS VALUE OF RESULT